

07.10.03

BUNDESREPUBLIK DEUTSCHLAND**PRIORITY
DOCUMENT**SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH RULE 17.1(a) OR (b)

RECD 13 OCT 2003
WIPO PCT

**Prioritätsbescheinigung über die Einreichung
einer Patentanmeldung**

18 03/4405

Aktenzeichen: 102 48 837.1

Anmeldetag: 19. Oktober 2002

Anmelder/Inhaber: Philips Intellectual Property & Standards GmbH,
Hamburg/DE
(vormals: Philips Corporate Intellectual Property GmbH)

Bezeichnung: System und Verfahren zur Verarbeitung von elektronischen Dokumenten

IPC: G 06 F 7/02

Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprünglichen Unterlagen dieser Patentanmeldung.

München, den 7. August 2003
Deutsches Patent- und Markenamt
Der Präsident
Im Auftrag

Klostermeyer

ZUSAMMENFASSUNG

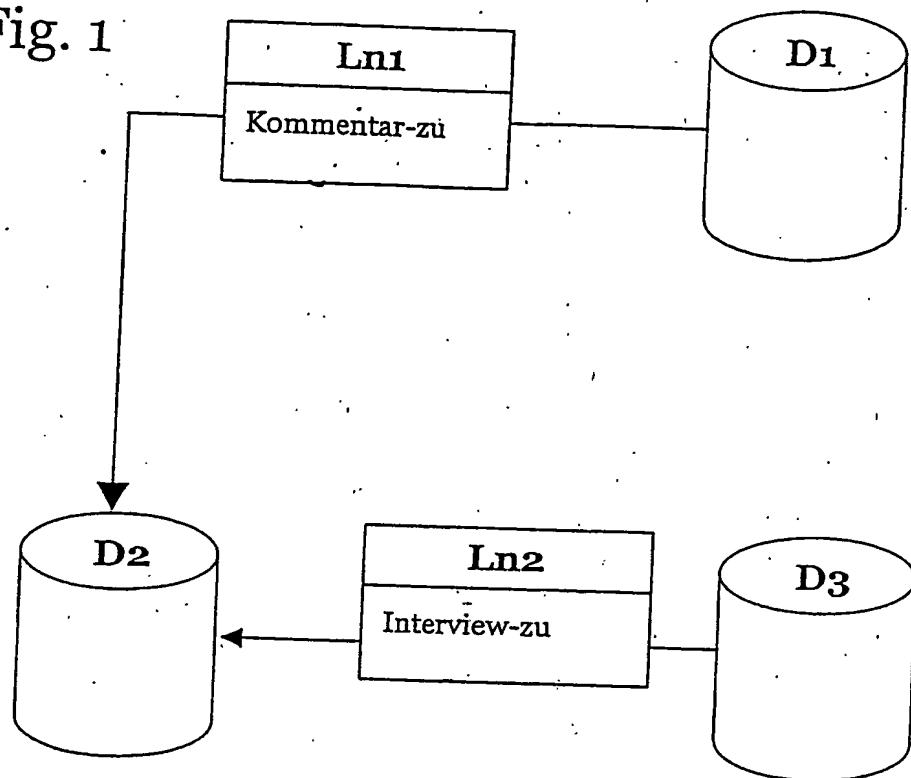
System und Verfahren zur Verarbeitung von elektronischen Dokumenten

- Ein System und ein Verfahren zur Verarbeitung von elektronischen Dokumenten werden beschrieben, bei denen ein Eingabedokument D1 und Referenzdaten D2
- 5 daraufhin untersucht werden, ob ein inhaltlicher Zusammenhang zwischen dem Eingabedokument D1 und den Referenzdaten D2 besteht. Für den Fall eines inhaltlichen Zusammenhangs wird ein Verknüpfungstyp entsprechend der Art des inhaltlichen Zusammenhangs aus einer Anzahl vorgegebener Verknüpfungstypen ausgewählt und eine entsprechende Verknüpfung zwischen den Dokumenten erstellt.
- 10 Die Erfindung ermöglicht, dass automatisch die Art der Beziehung zwischen zwei Dokumenten erkannt wird. So lässt sich bspw. ein Strom von Dokumenten in geeigneter Weise segmentieren und klassifizieren sowie sinnvoll vernetzt ablegen.

Fig. 1

PHDE020233

Fig. 1



BESCHREIBUNG

System und Verfahren zur Verarbeitung von elektronischen Dokumenten

5

Die Erfindung betrifft ein System und ein Verfahren zur Verarbeitung von elektronischen Dokumenten sowie ein Programm zur Durchführung des Verfahrens.

Angesichts der Vielzahl von heute zur Verfügung stehenden Daten, die bspw. über

10 Computernetzwerke wie das Internet abrufbar sind, wird verstärkt auf Systeme und Verfahren zurückgegriffen, die elektronische Dokumente entsprechend ihres Inhaltes automatisch verarbeiten. Bekannt sind hier bspw. Verfahren, die ein Dokument entsprechend seinem Inhalt klassifizieren.

15 In der US-A-5,983,246 sind ein Verfahren und eine Vorrichtung zur Verarbeitung von Dokumenten beschrieben. In einer Netzwerk-Umgebung werden ständig neue Dokumente bzw. neue Fassungen von Dokumenten aufgesucht und verarbeitet, indem sie nach ihrem Inhalt klassifiziert werden. Die Klassifizierung erfolgt automatisch, indem Ähnlichkeiten zwischen den aktuell bearbeiteten und bereits klassifizierten Dokumenten
20 ausgenutzt werden. Konkret wird ein Unterscheidungswert in Form einer Worthäufigkeits-Tabelle betrachtet, um ein Maß für die Übereinstimmung der Dokumente zu ermitteln.

Es ist Aufgabe der Erfindung, ein System und ein Verfahren anzugeben, mit dem Dokumente verarbeitet werden können und hierbei zusätzliche Informationen über die Dokumente automatisch generiert werden.

Diese Aufgabe wird gelöst durch ein System nach Anspruch 1, ein Verfahren nach Anspruch 11 und ein Programm nach Anspruch 12 zur Durchführung des Verfahrens. Abhängige Ansprüche beziehen sich auf vorteilhafte Ausführungsformen der Erfindung.

Erfnungsgemäß wird mindestens ein Eingabedokument im Hinblick auf einen inhaltlichen Zusammenhang mit Referenzdaten analysiert. Bei den Referenzdaten kann es sich bspw. um ein zweites Dokument handeln. Ebenso kann es sich bei den Referenzdaten um eine Gruppe (Cluster) von Dokumenten handeln oder um eine Repräsentation hierfür. Auf der Basis der Analyse wird entschieden, ob ein inhaltlicher Zusammenhang vorliegt. Dann wird die Art dieses Zusammenhangs ermittelt und versucht, diese einem Typ zuzuordnen. Hierfür sind eine Anzahl von möglichen Verknüpfungstypen, d.h. Arten von inhaltlichen Beziehungen zwischen zwei Dokumenten vorgegeben. Bei Vorliegen eines entsprechenden inhaltlichen Zusammenhangs wird eine entsprechende Verknüpfung zwischen den Dokumenten erstellt.

Unter "Dokumenten" werden hierbei Daten verstanden, die in elektronischer Form vorliegen. Es kann sich bspw. um Textdokumente handeln. Ebenso kann es sich um Kombinationen aus Text- und Bildinformationen handeln. Es wird bevorzugt, dass die verarbeiteten Dokumente mindestens einen Text-Teil aufweisen. Auch bspw. Audio- oder Videodateien können verarbeitet werden, wobei der Text-Inhalt dann bevorzugt entweder in transkribierter Form vorliegt oder auch bei der Verarbeitung durch ein Spracherkennungssystem generiert wird. Beispiele für Datei-Formate zu verarbeitender Dokumente sind HTML- oder – allgemeiner – XML-Dokumente. Die Dokumente können verschiedenen inhaltlichen Typs sein. Es kann sich bspw. um einzelne Nachrichten-Meldungen handeln. Ebenso können die Dokumente Werke der Literatur sein, oder wissenschaftliche Aufsätze, Interviews usw. Bevorzugt umfassen die Dokumente auch mindestens einen Daten-Teil mit zusätzlichen Informationen (Meta-Daten), z. B. eine Angabe der Quelle, ein Erstellungsdatum etc..

25

Im Rahmen der Erfindung sind eine Anzahl von Verknüpfungstypen vorgegeben. Diese Verknüpfungstypen entsprechen inhaltlichen Beziehungen zwischen zwei Dokumenten oder zwischen einem Dokument und einer Gruppe (Cluster) von Dokumenten. Beispiele für Verknüpfungstypen zwischen zwei Dokumenten A und B wären bspw. "Dokument

- A ist ein Interview zu dem in Dokument B geschilderten Ereignis" oder "Dokument A ist eine Rezension des Buches Dokument B". Entscheidend ist, dass ein inhaltlicher Zusammenhang besteht, der durch den Verknüpfungstyp festgelegt wird. Bevorzugt hat eine solche Verknüpfung eine festgelegte Richtung. Ein Beispiel für einen Cluster C wäre bspw. gegeben durch eine Gruppe von Dokumenten, die sich alle mit einem bestimmten Ereignis beschäftigen. Ein möglicher Verknüpfungstyp zwischen einem Dokument A und dem Cluster C wäre dann bspw. "Dokument A ist eine Diskussion über das Ereignis, von dem Cluster C handelt".
- 5 10 Die Erfindung geht somit über das bloße Feststellen von Ähnlichkeitsbeziehungen zwischen zwei Dokumenten hinaus. Automatisch wird die Art der Beziehung zwischen zwei Dokumenten oder einem Dokument und einem Cluster erkannt. So lässt sich bspw. ein Strom von Dokumenten in geeigneter Weise segmentieren und klassifizieren bzw. mit automatisch erzeugten Meta-Daten anreichern und sinnvoll vernetzt ablegen.
- 15 Das erfindungsgemäße System verfügt über Eingabemittel, Analysemittel, Auswahlmittel und Ausgabemittel. Bevorzugt handelt es sich um eine Vorrichtung mit einem oder mehreren Computern, die Dokumente und Referenzdaten bspw. aus einem Speicher oder über eine Netzwerkschnittstelle einlesen können. Die Analyse des
- 20 25 Zusammenhangs zwischen den Dokumenten und Referenzdaten sowie die Auswahl eines Verknüpfungstyps kann durch ein geeignetes Programm erfolgen. Die Ausgabe der erstellten Verknüpfung erfolgt bspw. durch Anzeigen auf einem Bildschirm, Ausgabe über eine Netzwerk-Schnittstelle oder Speicherung in einem geeigneten permanenten oder temporären Speicher.
- Gemäß einer Weiterbildung der Erfindung werden bei der Analyse der Dokumente Schlüsselworte aufgesucht, die die Art des Zusammenhangs zwischen den Inhalten des Eingabedokuments und der Referenzdaten bezeichnen. Entsprechend der aufgefundenen Schlüsselworte wird die Verknüpfung erstellt, d.h. der Verknüpfungstyp ausgewählt.

Beispiele für derartige Schlüsselworte können im Fall der Verarbeitung von Nachrichten-Dokumenten bspw. einleitende Worte sein wie "nun ein Kommentar zu ...". Bevorzugt handelt es sich um Kombinationen aus mehreren zusammenhängenden Schlüsselworten, die hier als Schlüsselphrasen bezeichnet werden.

5

Bei der Verarbeitung eines Dokuments kann dieses klassifiziert, d.h. zu einem von einer Anzahl vorgegebener Dokumenttypen zugeordnet werden. Die Bestimmung der Art des inhaltlichen Zusammenhangs kann dann auf den ermittelten Dokumenttyp zurückgreifen.

10.

Eine Weiterbildung der Erfindung sieht vor, dass das Eingabedokument einen Text-Teil und einen Daten-Teil umfasst. Der Text-Teil ist der bevorzugt verarbeitete Inhalt des Dokuments. Im Daten-Teil sind weitere Informationen (Meta-Daten) über das Dokument enthalten, bspw. Informationen über Art, Herkunft und/oder Datum des Dokumentes. Selbstverständlich kann das Dokument noch weitere Teile umfassen, bspw. Grafiken, Video- oder Audioinhalte. Die im Daten-Teil enthaltenen Meta-Daten über das Dokument können automatisch bei der Erfassung des Dokuments erstellt werden. Werden bspw. Nachrichtenbeiträge eines Fernsehsenders als Dokumente erfasst, so können die Quelle (Name des Nachrichtensenders) und die Sendezeit automatisch verzeichnet werden. Bei im Internet abgerufenen Dokumenten kann der Inhalte-Anbieter verzeichnet werden und, soweit abrufbar, weitere Meta-Daten (bspw. Erstellungsdatum, Name des Autors etc.). Weiter können Meta-Daten durch zusätzliche Verarbeitungsschritte generiert werden. Werden bspw. Dokumente verarbeitet, die ursprünglich als Audio- oder Videodateien vorlagen, und deren Textinhalt bspw. durch eine Spracherkennung generiert wird, so können weitere Informationen aus der Spracherkennung als Meta-Daten verarbeitet werden. Hierfür kann bspw. eine Identifikation des jeweiligen Sprechers vorgenommen werden. Derartige Techniken sind dem Fachmann aus dem Bereich der Spracherkennung bekannt. Die Ergebnisse der Sprecheridentifikation und bspw. auch ein regelmäßiger Sprecherwechsel (der auf den Dokumenttyp

"Interview" hindeuten würde) kann bspw. im Daten-Teil des Dokuments verzeichnet werden. Ebenso kann die Geräuschkulisse ausgewertet werden, um zwischen Studio-Beiträgen und bspw. Live-Reportagen (mit Hintergrundgeräuschen) zu unterscheiden und dies im Daten-Teil verzeichnet werden.

5

- Gemäß einer anderen Weiterbildung der Erfindung wird bei der Analyse des inhaltlichen Zusammenhangs der Dokumente auf eine spezielle Datenbank zugegriffen. In dieser Datenbank sind Begriffe der jeweiligen Sprache zugehörigen Oberbegriffen zugeordnet. Diese Informationen, angewendet auf Begriffe die in einem der beiden Dokumente vorkommen, können bei der Analyse des inhaltlichen Zusammenhangs zwischen den Dokumenten eingesetzt werden.

10

- Eine Weiterbildung der Erfindung betrifft die vernetzte Ablage von Dokumenten in einem elektronischen Speichersystem, in dem Dokumente semantisch vernetzt abgelegt sind. Zu abgespeicherten Dokumenten kann – wenn inhaltlich zugehörige Dokumente ebenfalls gespeichert sind – eine auf diese Dokumente gerichtete Verknüpfung des jeweiligen Verknüpfungstyps abgespeichert sein. Ein derartiges Speichersystem kann durch aufeinanderfolgende Verarbeitung von Dokumenten aufgebaut und um neue Dokumente erweitert werden. Beim Zugriff auf das Speichersystem kann zu einem Dokument auf einfache Weise, ohne zusätzliche Analyse-Schritte, auf inhaltlich zugehörige Dokumente zugegriffen werden. Über den Verknüpfungstyp kann der Zugriff gezielt auf bestimmte Arten von inhaltlichem Zusammenhang gerichtet werden. Das Speichersystem kann Teil des erfindungsgemäßen Computersystems sein und ein oder mehrere Speichermedien, bspw. elektronischen Speicher (RAM) und/oder optische bzw. magnetische Datenträger umfassen. Mehrere Speichermedien können zusammen in einem Gerät oder verteilt in mehreren, bspw. über ein Netzwerk miteinander verbundenen Geräten angeordnet sein.

15

20

25

Nachfolgend werden Ausführungsformen der Erfindung anhand von Zeichnungen näher beschrieben. In den Zeichnungen zeigen:

- Fig. 1: In symbolischer Darstellung Verknüpfungen zwischen drei Dokumenten;
- 5 Fig. 2: in symbolischer Darstellung Elemente eines Informationsverarbeitungssystems.

In Figur 1 sind in symbolischer Darstellung die drei Dokumente D1, D2 und D3 dargestellt.

- 10 Im vorliegenden Beispiel handelt es sich bei dem Dokument D2 um eine Video-Datei, die über ein aktuelles Ereignis berichtet. Die Videodatei ist Teil einer Nachrichtensendung und verfügt über einen Audio-Kommentar zum gezeigten Ereignis. Der Audio-Kommentar liegt in transkribierter Form zum Dokument D2 vor, bspw. erzeugt durch eine automatische Spracherkennung. Das Dokument D2 verfügt somit über einen
- 15 Video-Teil und einen Text-Teil. Zusätzlich verfügt das Dokument D2 über einen Daten-Teil, in dem Informationen über das Dokument gespeichert sind, darunter die ursprüngliche Sende-Zeit des Beitrags sowie die Bezeichnung des Senders.

- Das Dokument D1 ist im vorliegenden Fall ein Zeitungs-Kommentar zu dem aktuellen Ereignis, über das in D2 berichtet wird. Das Dokument D1 liegt in Form einer HTML-Seite mit dem entsprechenden Text vor. Zusätzlich zu dem Text-Teil verfügt auch D1 über einen Daten-Teil, in dem die Quelle (Name der Zeitung) sowie das Datum der Veröffentlichung verzeichnet sind.
- 25 Bei dem Dokument D3 handelt es sich um ein Interview zu demselben aktuellen Ereignis, von dem auch D2 handelt. Das Interview liegt als Audio-Datei vor. Mit Hilfe einer automatischen Spracherkennung wurde zudem der Wortlaut des Interviews in Textform umgewandelt, der so zur Verarbeitung zur Verfügung steht. Auch hier ist ein Daten-Teil mit Informationen über das Dokument vorhanden. Bei der Durchführung der

automatischen Spracherkennung wurde eine Sprecherkennung durchgeführt. Das erkannte Muster des regelmäßigen Wechsels zwischen zwei Sprechern (Interview) wurde erkannt und im Daten-Teil gespeichert.

- 5 Ein System zum Verarbeiten der Dokumente D1, D2 und D3 und zum Erzeugen von Verknüpfungen ist gegeben durch eine Datenquelle, die die Dokumente bereitstellt und durch einen Computer, der ein Programm verarbeitet, mit dem eine inhaltliche Beziehung zwischen zwei Dokumenten erkannt und eine entsprechende Verknüpfung zwischen den Dokumenten erstellt werden kann. Das Programm liest hierfür die Dokumente ein und verarbeitet den Text-Inhalt der Dokumente sowie ggfs. den Daten-Teil.
10 Hierbei wird zunächst festgestellt, ob inhaltliche Beziehungen zwischen den Dokumenten bestehen und welcher Art sie sind. Die Art der inhaltlichen Beziehung wird einer von einer vorgegebenen Liste von Verknüpfungsarten zugeordnet. Es wird eine Verknüpfung des ausgewählten Verknüpfungstyps zwischen den Dokumenten erzeugt.

- 15 Figur 1 zeigt eine Verknüpfung Ln1 zwischen den Dokumenten D1 und D2. Die Verknüpfung Ln1 ist vom Typ "Kommentar-zu". Die Verknüpfung ist gerichtet und zeigt von Dokument D1 auf Dokument D2. Sie gibt somit als inhaltlichen Zusammenhang zwischen D1 und D2 an, dass der Inhalt von D1 ein Kommentar ist zu dem in D2 geschilderten Ereignis.
20

- Ein anderes Beispiel ist eine Verknüpfung Ln2 zwischen den Dokumenten D3 und D2. Die Verknüpfung ist vom Typ "Interview-zu-Ereignis" und zeigt von Dokument D3 auf Dokument D2. Die Verknüpfung Ln2 wird von dem oben genannten Programm erzeugt
25 nachdem erkannt wurde, dass der Inhalt von D3 ein Interview zu dem im Dokument D2 geschilderten Ereignis ist.

Die in Fig. 1 dargestellten Dokumente D1, D2 und D3 mit den Verknüpfungen Ln1, Ln2 bilden eine Gruppe von Dokumenten, die hier als Cluster C bezeichnet wird. Ein solcher Cluster kann eine große Anzahl an Dokumenten umfassen. Die Dokumente eines Clusters hängen inhaltlich in der Weise zusammen, dass sie sich mit demselben
5 Thema befassen.

Die in Fig. 1 dargestellten Verknüpfungen Ln1 und Ln2 zwischen den Dokumenten D1, D2 und D3 sind jeweils Verknüpfungen zwischen einzelnen Dokumenten. Ebenso ist es auch möglich, Verknüpfungen zwischen einem neuen, zu analysierenden Dokument und
10 einem bestehenden Cluster C aus mehreren Dokumenten zu definieren.

Die Verarbeitung von Dokumenten durch das Programm läuft wie folgt ab:

- Zunächst wird ein Eingabedokument eingelesen. Bei der Bearbeitung wird einerseits der Text-Inhalt und andererseits ein Daten-Teil mit zusätzlichen
15 Informationen über das Dokument betrachtet.
- Das Eingabedokument wird mit Referenzdaten verglichen um festzustellen, ob ein
inhaltlicher Zusammenhang besteht. Wie oben erläutert kann es sich bei den
Referenzdaten um ein zweites Dokument handeln. Ebenso kann es sich bei den
20 Referenzdaten auch um einen Cluster von Dokumenten, bzw. um einen
Repräsentanten hiervon handeln.
- Wird keine inhaltliche Übereinstimmung zwischen dem Eingabedokument und
den Referenzdaten festgestellt, so ist die Verarbeitung hinsichtlich dieses
Vergleichspaares beendet. Das Eingabedokument kann dann bspw. mit weiteren
25 Referenzdaten verglichen werden.

- Wird hingegen ein inhaltlicher Zusammenhang festgestellt, erfolgt eine weitere Verarbeitung mit dem Ziel, die Art des Zusammenhangs zu ermitteln und eine entsprechende Verknüpfung zu generieren. Hierfür werden vordefinierte Schlüsselphrasen im Eingabedokument identifiziert, die einen Verweis aufeinander anzeigen. Den jeweiligen Schlüsselphrasen sind in einer Tabelle 5 Verknüpfungstypen zugeordnet.
 - Zusätzlich werden die im Daten-Teil des Eingabedokuments enthaltenen 10 Informationen ausgewertet. Die Ergebnisse der Schlüsselphrasen-Suche und die zusätzlichen Informationen aus dem Daten-Teil des Eingabedokuments werden bewertet, um einen Verknüpfungstyp auszuwählen.
- 15 Eine Verknüpfung des ausgewählten Verknüpfungstyps wird zwischen dem Eingabedokument und den Referenzdaten erzeugt und in einer Datenbank abgespeichert.
- Für die Feststellung, ob zwischen dem Eingabedokument und den Referenzdaten ein inhaltlicher Zusammenhang besteht, können dem Fachmann bekannte Techniken eingesetzt werden. Eine bekannte Technik umfasst eine Analyse des Text-Inhalts durch 20 Betrachtung häufig vorkommender Worte innerhalb des Textes. Werden zwei Dokumente verglichen, wird für beide Dokumente bspw. ein Vektor der Worthäufigkeiten der n häufigsten Worte erstellt, wobei n geeignet gewählt wird. Es kann dann ein Vektor-Abstand ermittelt werden, der als Maß für inhaltliche Übereinstimmungen zwischen den Dokumenten angesehen werden kann. Derartige Techniken sind bspw. in 25 der US-A-5 983 246 beschrieben. In den Artikeln "Text Categorization With Support Vector Machines: Learning with Many Relevant Features" 1998 by Thorsten Joachims, Proceedings of the ECML '98 (European Converence on Machine Learning) und "Improving text retrieval for the routing problem using latent semantic indexing" (1994) by David Hull, Proceedings of the SIGIR '94 (Special Interest Group on Information

(Retrieval) werden ebenfalls derartige Techniken diskutiert. Der Inhalt der zitierten Dokumente wird hier einbezogen.

Erfolgt eine Betrachtung des Zusammenhangs zwischen einem Dokument und einem Cluster von Dokumenten, so kann dies als Summe von Einzelvergleichen durchgeführt werden. Aus Performance-Gründen kann aber auch ein Vergleich des Dokuments mit einer oder mehreren Repräsentationen des Clusters erfolgen. Derartige Repräsentationen fassen Gemeinsamkeiten der Dokumente des Clusters zusammen. Wird bspw. mit der oben angegebenen Worthäufigkeit-Methode gearbeitet, so umfasst eine Repräsentation eines Clusters eine Liste von Begriffen, die in den Dokumenten des Clusters häufig vorkommen.

Der oben genannte Schritt der Auswahl eines geeigneten Verknüpfungstyps macht unter anderem Gebrauch von einer Tabelle mit Zuordnung von Schlüsselphrasen zu Verknüpfungstypen. Bei den Schlüsselphrasen kann es sich um einzelne Wörter handeln. In der Regel wird es sich jedoch um Kombinationen von Schlüsselworten und weiteren Elementen, wie Orts- oder Personennamen handeln. Nachfolgend ist beispielhaft eine Tabelle mit einer entsprechenden Zuordnung angegeben:

Schlüsselphrase	zugehöriger Verknüpfungstyp
Live vor Ort in <Ortsname> ist für uns <Personenname>	Live-Reportage
Dazu ein Kommentar von <Personenname>	Kommentar

20

Zusätzlich zu den oben angegebenen Schlüsselphrasen können Informationen mit Meta-Daten zum Eingabedokument verarbeitet werden. Derartige Meta-Daten können im Datenteil des Dokuments bereits enthalten sein, oder durch separater Verarbeitungsschritte generiert werden. So kann bspw. bei Erstellung des Text-Teils aus einer Audio-

- Datei zusätzlich zu bekannten Techniken der Spracherkennung auch die ebenfalls bekannten Techniken zur Sprecheridentifikation eingesetzt werden, um bspw. Regelmäßige Sprecherwechsel zu erkennen, die auf ein Interview hindeuten.
- 5 Die Gesamtheit der aus der Analyse der Schlüsselphrasen und der zusätzlichen Metadaten gewonnenen Informationen wird hinsichtlich der Übereinstimmung mit einem passenden Verknüpfungstyp bewertet. Der Verknüpfungstyp mit der höchsten Bewertung wird ausgewählt.
- 10 Zusätzlich kann bei der Analyse der Art der inhaltlichen Beziehung zwischen den Dokumenten auf eine spezielle Begriffs-Datenbank zugegriffen werden. Diese Datenbank enthält Begriffe der jeweils verwendeten Sprache und ordnet hierbei Begriffe einerseits ihren übergeordneten Oberbegriffen und andererseits von ihnen umfassten Spezialbegriffen zu. Das Wort "Werkzeug" wird so bspw. einerseits einem Oberbegriff "Gegenstand" zugeordnet und andererseits einem Spezialbegriff wie "Hammer". Derartige Datenbanken sind bekannt. Weiter verzeichnen bekannte Datenbanken dieser Art, die auch als "Thesaurus" bezeichnet werden, Synonyme und Antonyme von Begriffen ebenso wie Meronyme, Holonyme, Hyperonyme und Hyponyme von Begriffen.

- 15 20 Eine derartige Datenbank kann einerseits eingesetzt werden bei dem Schritt der Analyse, ob ein inhaltlicher Zusammenhang zwischen Eingabedokument und Referenzdaten besteht. Basiert diese Untersuchung auf einem Vergleich häufig auftretender Wörter, so können bspw. anstatt der Betrachtung von Einzelbegriffen Gruppen gleichbedeutender Begriffe (Synonyme) betrachtet werden, so dass unterschiedliche Formulierungen desselben Sachverhalts als inhaltlich zusammenhängend erkannt werden.

25 Andererseits können derartige Datenbanken auch bei der Feststellung der Art des inhaltlichen Zusammenhangs zwischen zwei Dokumenten bzw. zwischen einem Dokument- und einem Dokumenten-Cluster eingesetzt werden. Bspw. können in einer Datenbank

- mit Zuordnung von Spezial- und Oberbegriffen die in einem ersten Dokument auftretenden Begriffe hinsichtlich ihrer Stellung in der Datenbank (Oberbegriffe: allgemeiner; Spezialbegriffe: spezieller) betrachtet werden und so ein geeignetes, bspw. numerisches Maß für den Grad der Spezialisierung der verwendeten Begriffe gebildet
- 5 werden. Wird bspw. bei zwei inhaltlich als zusammenhängend erkannten Dokumenten festgestellt, dass ein Dokument überwiegend allgemeine Oberbegriffe nennt, während das andere Dokument Spezialvokabular verwendet, so können hieraus Rückschlüsse auf die unterschiedlich stark detaillierte Behandlung desselben Themas gezogen werden.
- 10 Diese Erkenntnisse können zusammen mit den Meta-Daten über das Dokument und Erkenntnissen über aufgefundene Schlüsselphrasen verwendet werden, um einen geeigneten Verknüpfungstyp auszuwählen.
- In Figur 2 ist in symbolischer Form ein System 10 zur Verarbeitung von Dokumenten 15 dargestellt. Das System 10 verfügt über einen Datenspeicher 12, in dem einerseits Dokumente D und andererseits Verknüpfungen L zwischen Dokumenten D abgelegt sind. Abgespeicherte, mit Verknüpfungen zusammenhängende Dokumente bilden Cluster C.
- Das System 10 verfügt ferner über eine Analyse- und Entscheidungseinheit 14 und eine 20 Auswahleinheit 16. Das System 10 verarbeitet ein Strom von Dokumenten D₁ ... D_n, die in ständiger Folge angeliefert werden. Dieser Strom von Dokumenten kann bspw. aus einer Dokumenten-Datenbank ausgelesen werden. Ebenso kann der Dokumentenstrom D₁ ... D_n das Ergebnis eines als "Web-Spider" arbeitenden Programms sein, das in ständiger Folge Dokumente aus dem Internet abruft. Der Datenstrom D₁ ... D_n kann 25 schließlich auch das Ergebnis einer ständigen Auswertung bspw. der Sendungen verschiedener Nachrichtensender sein.

Die Dokumente D₁ ... D_n werden zunächst von der Analyse- und Entscheidungseinheit 14 auf einen inhaltlichen Zusammenhang zu jedem der bereits im Datenspeicher 12 abgespeicherten Einzeldokumente D und Dokument-Clustern C überprüft. Bei Vorliegen einer inhaltlichen Beziehung wird wie oben angegeben deren Art ermittelt und 5 eine entsprechende Verknüpfung L erstellt. Das aktuell verarbeitete Dokument und sämtliche erzeugten Verknüpfungen L werden im Datenspeicher 12 abgelegt. So entsteht im Datenspeicher 12 ein semantisches Netzwerk, das Dokumente und gerichtete Relationen verschiedenen Typs zwischen diesen Dokumenten verzeichnet. Wird für ein Eingabe-Dokument kein Dokument D oder Cluster C mit inhaltlichem Zusammenhang 10 aufgefunden, so wird das Eingabedokument separat abgespeichert und kann den Kern eines neuen Referenz-Clusters bilden.

In einer konkreten Realisierung kann der Datenspeicher 12 bspw. als XML-Datenbank realisiert werden. Sind die Dokumente D bspw. in einem Computer-Netzwerk wie dem 15 Internet unter einer bekannten Adresse (URL) abrufbar, kann anstatt der Speicherung der Dokumente D im Datenspeicher 12 auch jeweils die entsprechende URL abgespeichert werden.

PATENTANSPRÜCHE

1. System zur Verarbeitung von elektronischen Dokumenten, mit
 - Eingabemitteln zur Eingabe mindestens eines Eingabedokuments (D1) und von Referenzdaten (D2)
- 5 - Analysemitteln (16) zur Analyse des Inhalts des Eingabedokuments (D1) hinsichtlich eines inhaltlichen Zusammenhangs zwischen dem Eingabedokument (D1) und den Referenzdaten (D2),
- 10 - Auswahlmitteln zur Auswahl eines Verknüpfungstyps aus einer Anzahl vorgegebener Verknüpfungstypen, wobei ein Verknüpfungstyp ausgewählt wird, entsprechend der Art des inhaltlichen Zusammenhangs zwischen dem Eingabedokument (D1) und den Referenzdaten (D2),
15 - und Ausgabemitteln zur Ausgabe einer Verknüpfung (L) des ausgewählten Typs.
2. System nach Anspruch 1, bei dem
 - die Verknüpfung (L) eine Verknüpfungsrichtung umfasst.
- 20 3. System nach einem der vorangehenden Ansprüche, bei dem
 - die Referenzdaten ein zweites Dokument (D2) sind.

4. System nach einem der Ansprüche 1 oder 2, bei dem

- die Referenzdaten eine Repräsentation für eine Gruppe von inhaltlich zusammenhängenden Dokumenten sind.

5. 5. System nach einem der vorangehenden Ansprüche, bei dem

- bei der Auswahl des Verknüpfungstyps Schlüsselworte aufgesucht werden, die die Art des Zusammenhangs zwischen den Inhalten des Eingabedokuments (D1) und der Referenzdaten (D2) bezeichnen,
- und ein Verknüpfungstyp entsprechend der aufgefundenen Schlüsselworte ausgewählt wird.

10

6. System nach einem der vorangehenden Ansprüche, bei dem

- bei der Auswahl des Verknüpfungstyps die Zuordnung des Dokuments (D) zu einem von einer Anzahl vorgegebener Dokumenttypen vorgenommen wird,
- und ein Verknüpfungstyp entsprechend des Dokumenttyps ausgewählt wird.

15

7. System nach einem der vorangehenden Ansprüche, bei dem

- das Eingabedokument (D1) mindestens einen Text-Teil und einen Daten-Teil umfasst,
- wobei der Daten-Teil Informationen enthält über die Art und/oder Herkunft des Dokuments.

20

8. System nach Anspruch 6 und 7, bei dem

- der Daten-Teil des Eingabedokuments (D1) zur Auswahl des Dokumenttyps verwendet wird.

25

9. System nach einem der vorangehenden Ansprüche, bei dem

- die Analysemittel auf eine Datenbank zugreifen, in der Begriffe zu Oberbegriffen zugeordnet sind.

5 10. System nach einem der vorangehenden Ansprüche, bei dem

- das Eingabedokument (D1) und die erstellte Verknüpfung (L) in einem Speichersystem (12) abgelegt wird,
- wobei das Speichersystem (12) so organisiert ist, dass zu darin gespeicherten Dokumenten jeweils Verknüpfungen zu anderen Dokumenten gespeichert sind.

10

11. Verfahren zur Verarbeitung von Dokumenten, bei dem

- mindestens ein Eingabedokument (D1) und Referenzdaten (D2) verarbeitet werden,
- wobei das Eingabedokument (D1) hinsichtlich seines Inhalts analysiert und entschieden wird, ob ein inhaltlicher Zusammenhang zwischen dem Eingabedokument (D1) und den Referenzdaten (D2) besteht,
- wobei für den Fall eines inhaltlichen Zusammenhangs ein Verknüpfungstyp aus einer Anzahl vorgegebener Verknüpfungstypen, entsprechend der Art des inhaltlichen Zusammenhangs zwischen dem Eingabedokument (D1) und den Referenzdaten (D2) ausgewählt wird,
- und eine Verknüpfung des ausgewählten Typs erstellt wird.

15

20

12. Programm zur Durchführung eines Verfahrens nach Anspruch 11.

25

PHDE020233

21

Fig. 1

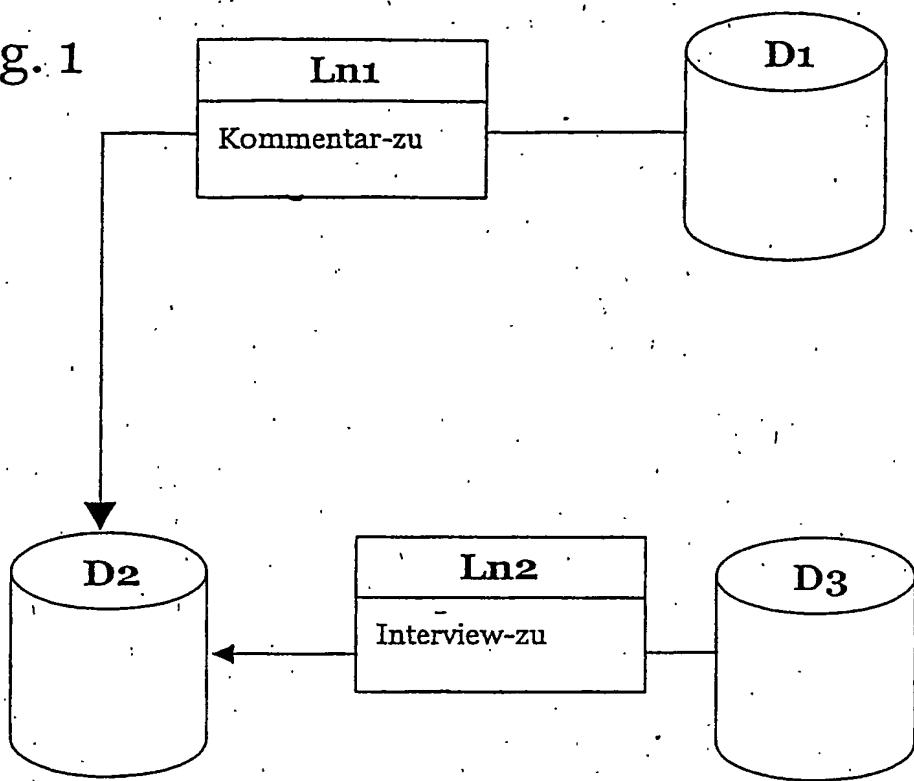
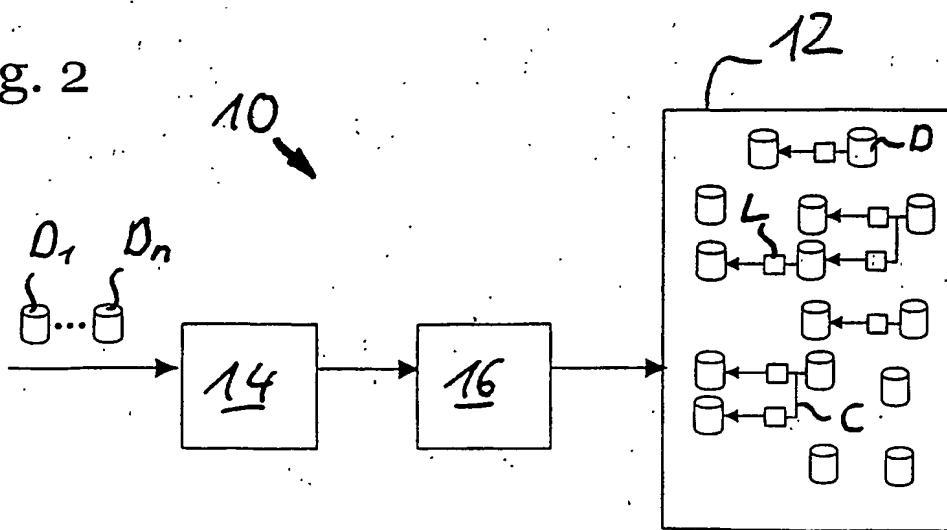


Fig. 2



**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- BLACK BORDERS**
- IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- FADED TEXT OR DRAWING**
- BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- SKEWED/SLANTED IMAGES**
- COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- GRAY SCALE DOCUMENTS**
- LINES OR MARKS ON ORIGINAL DOCUMENT**
- REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.